Today: Universal Models/Codes

1. normalized maximum likelihood code for countably infinite models (Chapter 7)
2. Bayesian marginal likelihood codes for countably infinite models (Chapter 8)
3. The amazing Jeffreys prior
   – Alternative for Laplace’s rule of succession
4. Questions/Feedback

Note: quite a lot of homework this time!

Universal Codes

- $\mathcal{L}$: set of code (length function)s available to encode data $x^n = (x_1, \ldots, x_n)$
- Suppose we think that one of the code(length function)s in $\mathcal{L}$ allows for substantial compression of x^n
- GOAL (for now): encode x^n using minimum number of bits!

Universal **Codes**

- Simply encoding x^n using the $\hat{L} \in \mathcal{L}$
  that minimizes code length
  does not work (encoding cannot be decoded)

- But there exist codes L which, for any sequence x^n are 'almost' as good as
  $\inf_{L \in \mathcal{L}} L(x^n)$

- These are called **universal codes** for $\mathcal{L}$

Regret Universal Model

Regret of distribution $\bar{p}$ on data x^n relative to model $\mathcal{M} = \{ p_\theta : \theta \in \Theta\}$ is given by:

$$-\log \bar{p}(x^n) - (-\log p_{\hat{\theta}(x^n)}(x^n))$$

Minimax Optimal Regret

$$\inf_{\bar{p}} \sup_{x^n \in \mathcal{X}^n} \left\{ -\log \bar{p}(x^n) - (-\log p_{\hat{\theta}(x^n)}(x^n)) \right\}$$

is achieved for Normalized Maximum Likelihood (NML) distribution (Shtarkov 1987):

$$\bar{p}_{\mathrm{nml}}(x^n) = \frac{p_{\hat{\theta}(x^n)}(x^n)}{\sum_{y^n \in \mathcal{X}^n} p_{\hat{\theta}(y^n)}(y^n)}$$

$$\inf_{\bar{p}} \sup_{x^n \in \mathcal{X}^n} \left\{ -\log \bar{p}(x^n) - (-\log p_{\hat{\theta}(x^n)}(x^n)) \right\}$$

is achieved for Normalized Maximum Likelihood (NML) distribution (Shtarkov 1987):

$$\bar{p}_{\text{nml}}(x^n) = \frac{p_{\hat{\theta}(x^n)}(x^n)}{\sum_{y^n \in \mathcal{X}^n} p_{\hat{\theta}(y^n)}(y^n)}$$

For all x^n, regret given by

$$-\log \bar{p}_{\text{nml}}(x^n) - [-\log p_{\hat{\theta}(x^n)}(x^n)] = \log \sum_{y^n \in \mathcal{X}^n} p_{\hat{\theta}(y^n)}(y^n)$$

(equalizer strategy)

How do the three Universal Codes Compare for finite model, $|\Theta| = K$?

- 2-part: worst-case regret bounded by $\log K$
- Bayes: worst-case regret (usually strictly) smaller

- NML: worst-case regret given by parametric complexity

$$\mathrm{comp}(\mathcal{M}) = \log \sum_{y^n \in \mathcal{X}^n} p_{\hat{\theta}(y^n)}(y^n)$$

- even (usually strictly) smaller

Parametric Complexity/ Minimax Regret, regular models

Finite $\mathcal{M}$:

$$\mathrm{comp}(\mathcal{M}) = \log\left(|\Theta| - \text{``total amount of confusion''}\right)$$

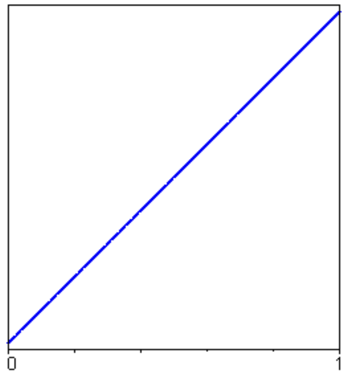
Countably infinite, “INECCSI” (≈ compact) Θ_0

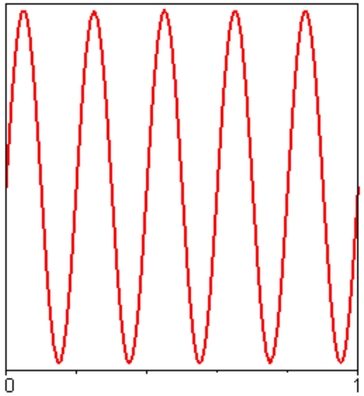
$$\mathrm{comp}(\mathcal{M}) = \frac{k}{2}\log\frac{n}{2\pi} + \log\int_{\Theta_0}\sqrt{\det I(\theta)} + o(1)$$

“geometric” contribution
to complexity/minimax regret

dimensional contribution
to complexity/minimax regret

Geometric Interpretation





Bernoulli vs. Crazy Bernoulli embedded in First-Order Markov

Regret of Bayes universal model

$$\bar{p}_{\mathsf{Bayes}}(x^n) := \int_{\Theta} p_\theta(x^n) w(\theta) d\theta$$

$$-\log \bar{p}_{\mathsf{Bayes}}(x^n) = -\log p_{\hat{\theta}(x^n)}(x^n) + \\ + \frac{k}{2} \log \frac{n}{2\pi} - \log w(\hat{\theta}(x^n)) + \log \sqrt{\det I(\hat{\theta}(x^n))} + o(1)$$

Regret of Bayes universal model

$$\bar{p}_{\text{Bayes}}(x^n) := \int_{\Theta} p_\theta(x^n) w(\theta) d\theta$$

$$-\log \bar{p}_{\text{Bayes}}(x^n) = -\log p_{\hat{\theta}(x^n)}(x^n) +$$
$$+\frac{k}{2}\log\frac{n}{2\pi} - \log w(\hat{\theta}(x^n)) + \log\sqrt{\det I(\hat{\theta}(x^n))} + o(1)$$

- convergence uniform for all x^n with $\hat{\theta}(x^n) \in \Theta_{\text{ineccsi}} \subset \Theta$ if prior continuous and bounded away from 0 on Θ_{ineccsi}
- within $O(1)$ of NML: for all 'reasonable' priors, Bayes gives universal model
- It can be better or worse than NML: **luckiness**

Regret of Bayes universal model

$$\bar{p}_{\text{Bayes}}(x^n) := \int_{\Theta} p_\theta(x^n) w(\theta) d\theta$$

$$-\log \bar{p}_{\text{Bayes}}(x^n) = -\log p_{\hat{\theta}(x^n)}(x^n) +$$
$$+\frac{k}{2}\log\frac{n}{2\pi} - \log w(\hat{\theta}(x^n)) + \log\sqrt{\det I(\hat{\theta}(x^n))} + o(1)$$

- convergence uniform for all x^n with $\hat{\theta}(x^n) \in \Theta_{\text{ineccsi}} \subset \Theta$ if prior continuous and bounded away from 0 on Θ_{ineccsi}
- within $O(1)$ of NML: for all 'reasonable' priors, Bayes gives universal model
- It can be better or worse than NML: **luckiness**
- **b**ut can it be made to **mimic NML?**

The Amazing Jeffreys' Prior

- In 1946, Sir Harold Jeffreys (who discovered that the interior of the earth is fluid) proposed what is now called Jeffreys' prior,

$$w(\theta) = \frac{\sqrt{\det I(\theta)}}{\int_{\Theta_0} \sqrt{\det I(\theta)} d\theta}$$

- …to be used "when real prior knowledge is lacking"

Regret of Bayes-Jeffreys

$$-\log \bar{p}_{\text{Bayes}}(x^n) = -\log p_{\hat{\theta}(x^n)}(x^n) +$$

$$+\frac{k}{2}\log\frac{n}{2\pi} - \log w(\hat{\theta}(x^n)) + \log\sqrt{\det I(\hat{\theta}(x^n))} + o(1)$$

- within $O(1)$ of NML: for all 'reasonable' priors, Bayes gives universal model
- But if we plug in Jeffreys' prior, within $o(1)$.
- **With Jeffreys prior, asymptotically Bayes and NML coincide!**

$$w_{\text{Jeffreys}}(\theta) = \frac{\sqrt{\det I(\theta)}}{\int_{\Theta_0}\sqrt{\det I(\theta)}d\theta}$$

More on Jeffreys' Prior

- $\bar{p}_{\text{B-J}}(x^n) := \int_{\Theta_0} p_\theta(x^n) w_{\text{Jeffreys}}(\theta) d\theta$
- often easier to compute than $\bar{p}_{\text{nml}}$
- …has been advocated as prior for model selection in the Bayesian literature – makes MDL and Bayes "consistent"

Jeffreys’ Prior vs Luckiness

More on Jeffreys' Prior

- Jeffreys' introduced his prior for different reasons
- Important Reason: invariance to **reparameterization** parameter space
  - in uncountable spaces, the notion of 'uniform' prior depends on choice of parameterization (and is hence arbitrary)
  - Example: Bernoulli can also be parameterized by $p_\theta(X = 1) = \theta^{\mathbf{2}}$. Uniform density on θ gives a very different distribution on the set of Bernoulli distributions than uniform density on θ in standard parameterization

More on Jeffreys' Prior

- Example: Bernoulli can also be parameterized by $p_\theta(X = 1) = \theta^{\mathbf{2}}$. Uniform density on θ gives a very different distribution on the set of Bernoulli distributions than uniform density on θ in standard parameterization
- Jeffreys' prior is **parameterization invariant**. (Hence a better choice for ignorance than Laplace-Bayes' choice, which was the uniform prior)

More on Jeffreys' Prior

- Jeffreys prior for Bernoulli:

$$\bar{p}_{\text{B-J}}(X_{n+1} = 1 \mid x^n) = \frac{n_1 + 1/2}{n+1}$$

- Jeffreys prior for Gaussian location:
  - uniform on Θ_0 (space of means)

  (parameter space must be restricted, otherwise prior "improper" – i.e. does not integrate)

"Luckiness Again"

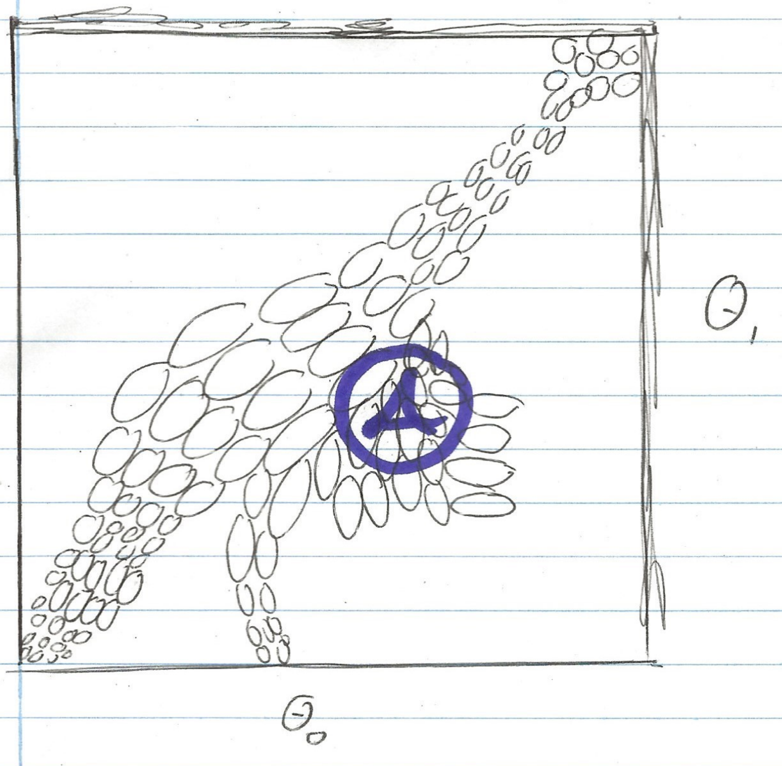
- See drawing: for what sequences does Jeffreys' prior lead to smaller regret, for what sequences to larger regret?

Geometric Interpretation of Jeffreys' prior

- Jeffreys' prior is **uniform prior on space of distributions rather than parameters…**
- …when 'distance' between distributions is measured by
  - KL divergence
  - 'distinguishability'

Next Week

- Simple MDL/Bayesian Model Selection:
  - again, almost the same!
- Yet another universal code/model: “prequential”



COVER PARAMETER SPACE

BY ε-KULLBACK LEIBLER BALLS

$$B_\Theta(\varepsilon) = \{\Theta' : D(\Theta \| \Theta') \leq \varepsilon\}$$

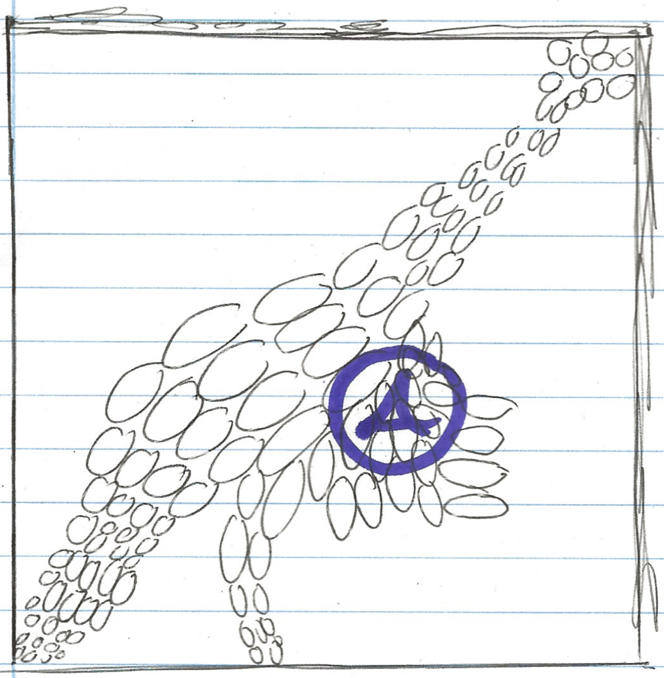
NOT METRIC

METRIC

$$\approx \{\Theta' : \tfrac{1}{2}(\Theta-\Theta')^T I(\Theta)(\Theta-\Theta') \leq \varepsilon\}$$

$\hookrightarrow$ ELLIPSOIDS IN PARAMETER SPACE

$$W_{JEFFREY}(A) = \lim_{\varepsilon \downarrow 0} \frac{\#\text{BALLS WITH RADIUS } \varepsilon \text{ IN } A}{\#\text{BALLS } \text{''} \quad \text{''} \quad \varepsilon \text{ IN PARAMETER SPACE}}$$



BALASUBRAMANIAN:

$P_\Theta(\hat{\Theta} \in B_\Theta(\varepsilon))$

$\approx$ $1 - C \cdot \varepsilon$

DISTINGUISHABILITY

COVER PARAMETER SPACE

BY ε-KULLBACK LEIBLER BALLS

$$B^{KL}_\Theta(\varepsilon) = \{\Theta' : D(\Theta \| \Theta') \leq \varepsilon\}$$

NOT METRIC

METRIC

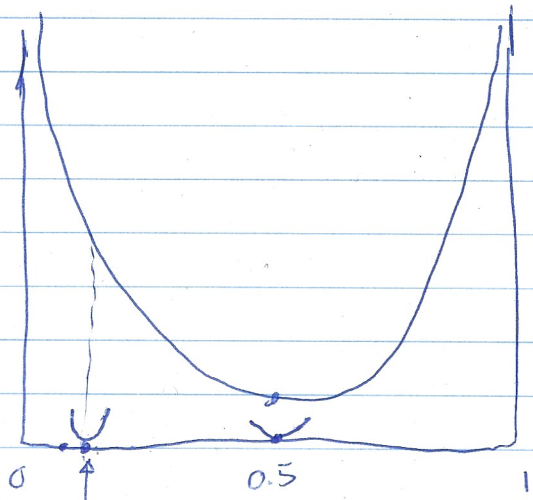
$$\approx \{\Theta' : \tfrac{1}{2}(\Theta - \Theta')^T I(\Theta)(\Theta - \Theta') \leq \varepsilon\}$$

$\hookrightarrow$ ELLIPSOIDS IN PARAMETER SPACE

$$W_{JEFFREY}(A) = \lim_{\varepsilon \downarrow 0} \frac{\#\text{BALLS WITH RADIUS } \varepsilon \text{ IN } A}{\#\text{BALLS " " } \varepsilon \text{ IN PARAMETER SPACE}}$$

$$D(\theta_0 \| \theta) \approx \frac{1}{2} I(\theta)(\theta_0 - \theta)^2$$

EXAMPLE BERNOULLI



$I(\theta)$

$\theta = 0.5: \; D(\theta \| \theta + \varepsilon) \approx \frac{1}{4} \varepsilon^2.$

$\theta = 0.1: \; D(\theta \| \theta + \varepsilon) \approx \frac{10}{0.9} \varepsilon^2 \approx 11 \varepsilon^2$

NORMAL: $I(\theta)$ CONSTANT

LOCATION

$$D(\theta \| \theta') = \frac{1}{2\sigma^2}(\theta - \theta')^2$$

"NO CURVATURE"

NEXT WEEK

TWO PART

- DISCRETIZE WITH WIDTH $\propto \sqrt{I(\theta)}$

  $\sqrt{\det I(\theta)}$

- BAYES: YOU USE PRIOR $\propto \sqrt{I(\theta)}$

- NML: ALSO RELATED TO $I(\theta)$.